iNeuron

HIGH LEVEL DESIGN

CROP PRODUCTION OF INDIA

Chetan Patil, Mohit Kumar Tomar, Mr Pushpendra Dudi, Sayani Mitra, Subhangi Doye

HIGH LEVEL DOCUMENTATION

Contents

HIGH LEVEL DOCUMENTATION

Abstract

Agriculture, with its allied sectors, is unquestionably the largest livelihood provider in India, more so in the vast rural areas. It also contributes a significant figure to the Gross Domestic Product (GDP).About 70% of the Indian population practices agriculture. Hence, the production and management of crops is an important aspect to ensure optimal productivity in the fields.

When plants of the same variety are cultivated on a large scale, they are called crops. The crops are divided on the basis of the seasons in which they grow: **Kharif Crops, Rabi Crops and Whole year Crops.**

Kharif crops

Kharif crops are typically sown at the beginning of the first monsoon rains (depending on region to region). Harvesting season begins from the 3rd week of September to October (the exact harvesting dates differ from region to region). Paddy, maize, bajra, jowar are a few of the Kharif crops grown in India.

Rabi Crops

Rabi crops are known as winter crops. They are grown in October or November. The crops are then harvested in spring. These crops require frequent irrigation because they are grown in dry areas. Wheat, gram, and barley are some of the Rabi crops grown in India.

Whole Year crop

The **whole year crop** definition is a plant that completes its entire life cycle in one year or one growing season. An annual crop, or yearly crop or plant, only lives for one growing season, which is the length of months that provides optimal growing conditions for that particular plant. Within the course of a single year, annual plants germinate, grow to maturity, and produce their own flowers, seeds,

and fruit before dying. Pulses, sugarcanes, potato, dry chillies etc are known as whole year crops or plants.

1. Introduction

1.1 Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

- Present all of the design aspects and define them in detail
- Describe the user interface being implemented
- Describe the hardware and software interfaces
- Describe the performance requirements
- Include design features and the architecture of the project
- List and describe the non-functional attributes like:
  - Security
  - Reliability
  - Maintainability
  - Portability
  - Reusability
  - Application compatibility
  - Resource utilization
  - Serviceability

1.2 Scope

HIGH LEVEL DOCUMENTATION

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

2 General Descriptions

2.1 Product Perspective

Agriculture plays a vital role in the Indian economy. It is the backbone of the country. Most of its population (about 60 per cent) is depends on Agriculture. Nearly 16 per cent of overall GDP of India (approx. Rs 19.48 lakh core or US$ 276.37 billion) is contributed by agriculture, forestry and fisheries. The weather conditions and soils conditions allow for crops in India is perfectly suited for it.

Problem Statement:

The Agriculture business domain, as a vital part of the overall supply chain, is expected to highly evolve in the upcoming years via the developments, which are taking place on the side of the Future Internet. This paper presents a novel Business-to-Business collaboration platform from the agri-food sector perspective, which aims to facilitate the collaboration of numerous stakeholders belonging to associated business domains, in an effective and flexible manner. This dataset provides a huge amount of information on crop production in India ranging from several years. Based on the Information the ultimate goal would be to predict crop production and find important insights highlighting key indicators and metrics that influence the crop production. Make views and dashboards first. Make a story out of it.

2.2 Tools used

Business Intelligence tools and python libraries works such as numpy, Pandas, Excel and Power BI are used to build the whole framework.



PREPROCESSING

Dataset

For this research, we have used **'Crop production of India'** dataset which has been provided by '**Ineuron.ai**'.This dataset contains more than 2, 40,000 data rows with detailed information of crop production in India from 1997 to 2015 with respect to different districts and States.
A lot of pre-processing was required to handle missing values, noise and outliers. We have considered 7different attributes for this research: ***State_name, District_name, Crop year, Season, Crop, Area and Production.***

METHODOLOGY

First we imported necessary libraries of python which we needed to precede our work.

```
import pandas as pd,os
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
import warnings
warnings.filterwarnings('ignore')
```

HIGH LEVEL DOCUMENTATION

After loading our data set we checked is there any null value present on any attributes and we discovered that for production attribute there were 3730 null values.

```
crop_df.isnull().sum()

State_Name 0
District_Name 0
Crop_Year 0
Season 0
Crop 0
Area 0
Production 3730
dtype: int64
```

We have approx 243000 rows and out of that 3730were null (1.53%) so we decided to drop these rows for our better analysis.

```
crop_df.dropna(inplace=True)
```

Now no null values were there in our dataset.

```
crop_df.isnull().sum()

State Name 0
District_Name 0
Crop_Year 0
Season 0
Crop 0
Area 0
Production 0
```

We started our analysis with season column.

Season

```
crop_df.Season.unique()

array(['Kharif', 'Whole Year', 'Autumn', 'Rabi', 'Summer', 'Winter'],
dtype=object)

# checked the values counts of each season
```

HIGH LEVEL DOCUMENTATION

```
crop_df['Season'].value_counts()
```

```
Kharif 94283
Rabi 66160
Whole Year 56127
Summer 14811
Winter 6050
Autumn 4930
```

Since there are three types of crop out of which two are mainly Seasonal Rabi and Kharif and another one is whole year but, 5 types of seasons were present in our dataset.
We got to know that summer and autumns are the synonyms of Kharif and Winter is the synonyms of Rabi so we decided to replace with their original name which is Kharif and Rabi.

```
crop_df['Season']=crop_df['Season'].apply(lambda x : x.replace('Autumn','Kharif'))
crop_df['Season']=crop_df['Season'].apply(lambda x : x.replace('Summer','Kharif'))
crop_df['Season']=crop_df['Season'].apply(lambda x : x.replace('Winter','Rabi'))
```

Now we had three types of seasons in our dataset that is Kharif, Rabi and whole year.

Production

We have observed that many values of production were 0(zero)
Since it is representing production of whole district
so we decided drop all rows whose production values are zero.

```
crop_df.drop(crop_df[crop_df['Production']==0].index,inplace=True)
print(f'After removing the row which has 0 Production : {crop_df.shape[0]}')
```

Crop

HIGH LEVEL DOCUMENTATION

We have observed that many crops were presented with their synonyms so we decided to replace all synonyms of crops with their Actual name like paddy is the synonyms of rice so we replaced paddy with rice.

```
crop_df.Crop.unique()

array(['Arecanut', 'Other Kharif pulses', 'Rice', 'Banana', 'Cashewnut', 'Coconut ',
'Dry ginger', 'Sugarcane', 'Sweet potato', 'Tapioca', 'Black pepper', 'Dry chillies',
'other oilseeds', 'Turmeric', 'Maize', 'Moong(Green Gram)', 'Urad', 'Arhar/Tur',
'Groundnut', 'Sunflower', 'Bajra', 'Castor seed', 'Cotton(lint)', 'Horse-gram', 'Jowar',
'Korra', 'Ragi', 'Tobacco', 'Gram', 'Wheat', 'Masoor', 'Sesamum', 'Linseed',
'Safflower', 'Onion', 'other misc. pulses', 'Samai', 'Small millets', 'Coriander',
'Potato', 'Other Rabi pulses', 'Beans & Mutter(Vegetable)', 'Bhindi', 'Brinjal', 'Citrus
Fruit', 'Grapes', 'Mango', 'Orange', 'Other Fresh Fruits', 'Papaya', 'Pome Fruit',
'Tomato', 'Soyabean', 'Mesta', 'Cowpea(Lobia)', 'Lemon', 'Pome Granet', 'Sapota',
'Cabbage', 'Rapeseed &Mustard', 'Niger seed', 'Varagu', 'Garlic', 'Ginger', 'Oilseeds
total', 'Pulses total', 'Jute', 'Peas & beans (Pulses)', 'Blackgram', 'Paddy',
'Pineapple', 'Barley', 'Sannhamp', 'Khesari', 'Guar seed', 'Other Vegetables', 'Moth',
'Other Cereals & Millets', 'Cond-spcs other', 'Turnip', 'Carrot', 'Redish', 'Arcanut
(Processed)', 'Atcanut (Raw)', 'Cashewnut Processed', 'Cashewnut Raw', 'Cardamom',
'Rubber', 'Drum Stick', 'Jack Fruit', 'Tea', 'Coffee', 'Total foodgrain', 'Cauliflower',
'Bitter Gourd', 'Bottle Gourd', 'Kapas', 'Colocosia', 'Lentil', 'Bean', 'Jobster',
'Perilla', 'RajmashKholar', 'Ricebean (nagadal)', 'Jute &mesta'],
```

We have replaced **Kapas** with **cotton** and **Jute and Mesta** with **Jute.**

```
crop_df['Crop']=crop_df['Crop'].apply(lambda x:x.replace('Kapas','Cotton(li
nt)')
                                      .replace('Cotton(lint)','Cotton'))

crop_df.replace('Jute & mesta','Jute',inplace=True)
crop_df.replace('Mesta','Jute',inplace=True)
```

We have replaced all the sub-category of pulses with pulse.

```
crop_df['Crop']=crop_df['Crop'].apply(lambda x:x.replace('Other Kharif puls
es','pulses')
                                    .replace('Other  Rabi pulses','pulses')

                                    .replace('Peas & beans (Pulses)','puls)
                                    .replace('Pulses total','pulses')
                                    .replace('other misc. pulses','pulses')
                                    .replace('Moong(Green Gram)','pulses')
                                    .replace('Urad','pulses')
```

```
                                    .replace('Arhar/Tur','pulses')
                                    .replace('Bean','pulses')
                                    .replace('Ricebean (nagadal)','pulses')
                                    .replace('Lentil','pulses')
                                    .replace('Masoor','pulses')
                                    .replace('Khesari','pulses')
                                    .replace('Horse-gram','pulses')
                                    .replace('Rajmash Kholar','pulses'))
```

We replaced all kind of spices with the name of other spices.

```
crop_df['Crop']=crop_df['Crop'].apply(lambda x:x
                                    .replace('Black pepper','Other Spices')
                                    .replace('Cardamom','Other Spices')
                                    .replace('Perilla','Other Spices'))
```

Since number of rows for every fruits category were very less so we have decided to merge all to fruits category with the name Fruits.

```
crop_df['Crop']=crop_df['Crop'].apply(lambda x:x
                                    .replace('Papaya','Fruits')
                                    .replace('Mango','Fruits')
                                     .replace('Orange','Fruits')
                                     .replace('Other Fresh Fruits','Fruits'
)
                                     .replace('Pineapple','Fruits')
                                     .replace('Citrus Fruit','Fruits')
                                     .replace('Pome Fruit','Fruits')
                                     .replace('Pome Granet','Fruits')
                                     .replace('Grapes','Fruits')
                                     .replace('Jack Fruit','Fruits')
                                     .replace('Sapota','Fruits')
                                     .replace('Lemon','Fruits'))
```

We replaced Ginger with Dry ginger, Turnip with Onion, Cashew nut Raw and Cashew nut processed to Cashew nut.

```
crop_df['Crop']=crop_df['Crop'].apply(lambda x:x
                                    .replace('Ginger','Dry ginger'))

crop_df['Crop']=crop_df['Crop'].apply(lambda x:x
                                    .replace('Turnip','Onion'))

crop_df['Crop']=crop_df['Crop'].apply(lambda x:x
                                     .replace('Cashewnut Raw','Cashewnut')
```

```
                                        .replace('Cashewnut Processed','Cashe
wnut'))
```

We also replaced Black gram and Moth with Gram

```
crop df['Crop']=crop df['Crop'].apply(lambda x:x
                                        .replace('black gram','gram')
                                       .replace('Moth','gram')
                                       .replace('Blackgram','gram'))
```

We replaced oilseeds total, Niger seed to other oilseeds

```
crop_df['Crop']=crop_df['Crop'].apply(lambda x:x
                                        .replace('Oilseeds total' ,'other oil
seeds')
                                        .replace('Niger seed' ,'other oilseed
s'))
```

We replaced sun-category of millets with their main category

```
crop_df['Crop']=crop_df['Crop'].apply(lambda x:x
                                        .replace('Other Cereals & Millets' ,'
Bajra')
                                        .replace('Samai' ,'Bajra')
                                        .replace('Small millets' ,'Bajra')
                                        .replace('Ragi','Bajra')
                                        .replace('Varagu','Bajra')
                                        .replace('Jobster','Bajra'))
```

All vegetables name with other vegetables.

```
crop_df['Crop']=crop_df['Crop'].apply(lambda x:x
                                        .replace('Coriander' ,'Other Vegetabl
es')
                                        .replace('pulsess & Mutter(Vegetable)
' ,'Other Vegetables')
                                        .replace('Bhindi' ,'Other Vegetables'
)
                                        .replace('Tomato','Other Vegetables')
                                        .replace('Cowpea(Lobia)','Other Veget
ables')
                                        .replace('Cabbage','Other Vegetables'
)
                                        .replace('Carrot','Other Vegetables')
```

HIGH LEVEL DOCUMENTATION

At last we have decided to drop tea, coffee and rubber as their production is showing very less in this dataset which is impossible because India is among the top 10 coffee-producing countries. Indian coffees and teas are one of the best in the world due to its high quality and get a high premium in the international markets.

```
crop_df.drop(crop_df[crop_df['Crop']=='Tea'].index,inplace=True)
crop_df.drop(crop_df[crop_df['Crop']=='Coffee'].index,inplace=True)
crop_df.drop(crop_df[crop_df['Crop']=='Rubber'].index,inplace=True)
crop_df.drop(crop_df[crop_df['Crop']=='Cond-
spcs other'].index,inplace=True)
```

After replacing all the names of crops the crop attribute now have the categories like-

```
crop_df.Crop.unique()

array(['Arecanut', 'pulses', 'Rice', 'Banana', 'Cashewnut', 'Coconut ',
'Dry ginger', 'Sugarcane', 'Sweet potato', 'Tapioca', 'Other Spices', 'Dry
chillies', 'other oilseeds', 'Turmeric', 'Maize', 'Groundnut', 'Sunflower',
'Bajra', 'Castor seed', 'Cotton', 'Jowar', 'Total foodgrain', 'Tobacco',
'Gram', 'Wheat', 'Sesamum', 'Linseed', 'Safflower', 'Onion', 'Other
Vegetables', 'Potato', 'Fruits', 'Soyabean', 'Jute', 'Rapeseed &Mustard',
'Garlic', 'gram', 'Barley', 'Sannhamp', 'Guar seed'], dtype=object)
```

We calculated production per area and created a new column “Production_area_factor” to keep these values.

```
crop_df['Producation_area_factor']= crop_df['Production'] / crop_df['Area']
```

India is a very big country, so for better understanding and visualization we created a zone column and we divided our states into 5 different Zones – West India, East India, North India, South India and Union Territory.

```
West_India= ['Maharashtra','Goa','Gujarat','Dadra and Nagar Haveli']
East_India= ['Arunachal Pradesh','Assam','Manipur','Meghalaya','Mizoram','Nagaland'
,'Sikkim','Tripura','West Bengal','Bihar','Odisha','Jharkhand' ]
North_India=['Jammu and Kashmir ','Himachal Pradesh','Punjab','Uttarakhand','Haryan
a','Rajasthan','Uttar Pradesh','Chandigarh','Madhya Pradesh','Chhattisgarh']   # De
lhi
South_India = ['Andhra Pradesh','Karnataka','Kerala','Tamil Nadu','Telangan
a','Puducherry'
```

HIGH LEVEL DOCUMENTATION

```
zone = []

for df in crop_df['State_Name']:
    if df in West_India:
         zone.append('West India')
    elif df in East_India:
         zone.append('EastIndia')
    elif df in North_India:
         zone.append('North India')
    elif df in South_India:
         zone.append('South India')
    else:
         zone.append('Union Territory')

crop_df['zone'] = zone
```

Count of states in different zones are

```
North India 92752
EastIndia 71482
South India 47554
West India 21229
Union Territory 5706
```

After that we checked value counts of year column, and we found that for the year 2015 data count were very low, so we dropped year 2015 from our data set.

```
crop_df['Crop_Year'].value_counts()
2003 15541
2002 15060
2007 14261
2008 14230
2006 13976
2004 13834
2010 13793
2011 13791
2009 13767
2005 13519
2013 13474
2000 13393
2012 13183
2001 13107
1999 12258
1998 11262
2014 10814
1997 8899
```

```
crop_df.drop(crop_df[crop_df['Crop_Year'] == 2015].index,inplace=True
```

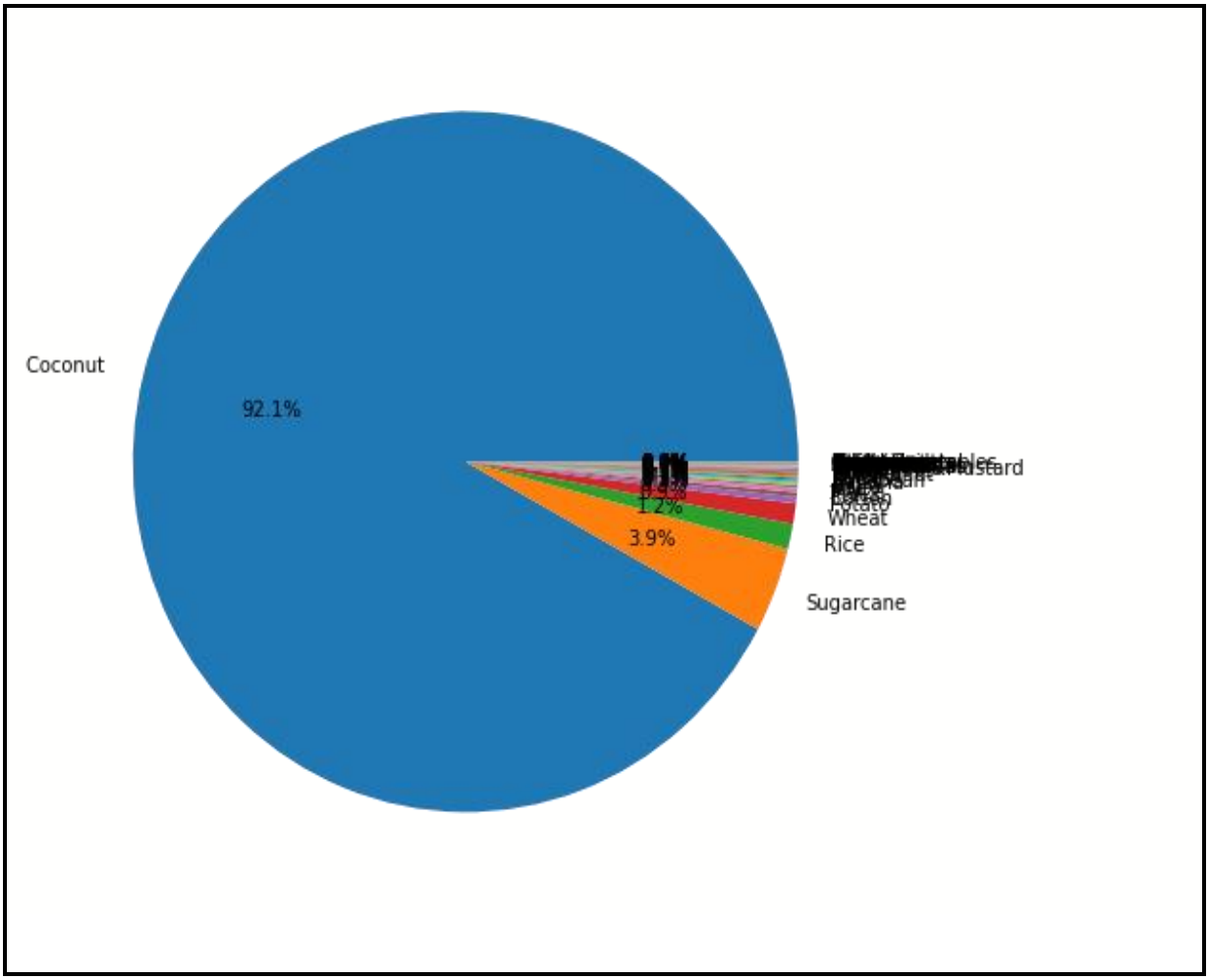
We plotted pie chart to check outliers.

```
val=crop_df.groupby('Crop').sum().sort_values(by='Production',ascending = F
alse)['Production'].values
```

HIGH LEVEL DOCUMENTATION

```
lab=crop_df.groupby('Crop').sum().sort_values(by='Production',ascending = F
alse)['Production'].index

plt.figure(figsize=(10,8))
plt.pie(val,labels=lab,autopct='%0.1f%%')
plt.show()
```



From the above chart we have observed that Production of coconut is more than 92% of total Production so we were not able to visualize our dataset because it behaves like an outlier. But on the other hand coconut is very important and productive crops of our India, so we can't ignore this crop.
So we have decided to visualize coconut separately, we have created 2 datasets for our visualization, one is with coconut data, and another one is without coconut data.

```
coconut_df = crop_df[crop_df['Crop'] == 'Coconut']
```

HIGH LEVEL DOCUMENTATION

```
coconut_df.to_csv('coconut_df.csv')
```

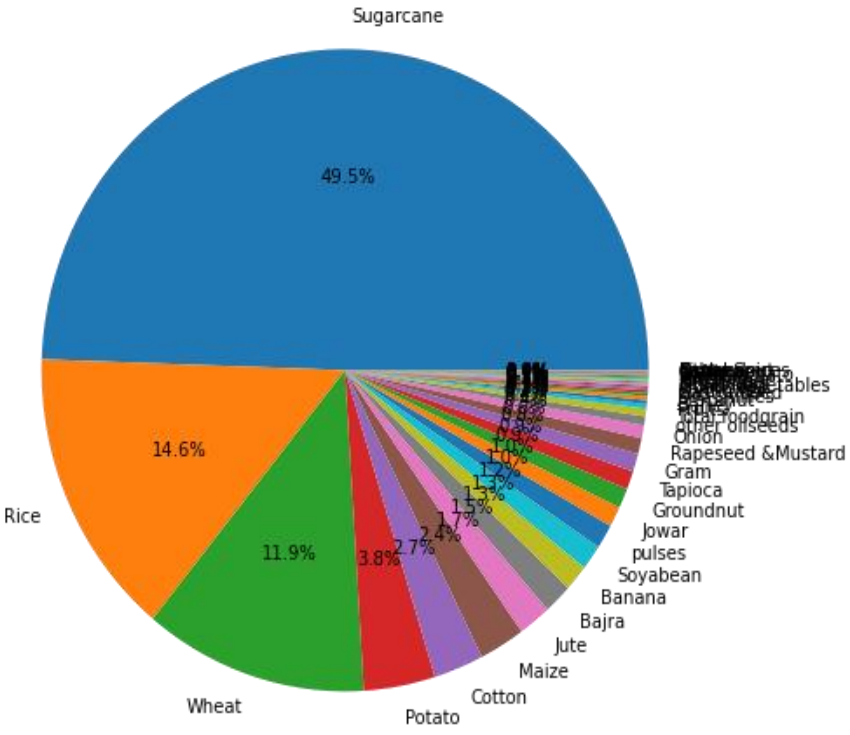
Removed coconut from our dataset.

```
crop_df.drop(crop_df[crop_df['Crop']=='Coconut'].index,inplace=True)
```

After that we plotted pie chart between total Production and with different types of crops.

```
val=crop_df.groupby('Crop').sum().sort_values(by='Production',ascending
 = False)['Production'].values
lab=crop_df.groupby('Crop').sum().sort_values(by='Production',ascending
 = False)['Production'].index
```

```
plt.figure(figsize=(10,8))
plt.pie(val,labels=lab,autopct='%0.1f%%')
plt.show()
```



Now our dataset became a balanced dataset, so that we can load these two datasets into Power BI and can start doing our visualization.

Conclusion

We used two different datasets for our visualization one with coconut values and another dataset for which we removed coconut values from there.

After this we used these datasets in Power BI and found different valuable insights from it which helped us to understand the variation in crop production of India based on different seasons, years, states and districts.